Self-Directed Lifelong Visual Learning

Rob Fergus NYU
Kristen Grauman TEXAS The University of Texas at Austin
Erik Learned-Miller UMass Amherst
Greg Shakhnarovich TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Driving Applications: Lifelong Learning Machines (L2M)

- Today's learning paradigms are ***stagnant*** : short-term, non-adaptive.
  - Over-reliant on labeled-data.
  - Can't apply past knowledge to new domains.
  - Can't start learning until task is presented.

Our approach: *Lifelong visual learning,*

- Continual, adaptive learning.
- Adjust to new domains, new tasks, new environments.
- Leverage massive unlabeled data sets of images/video.
- Learn to see and act without labels via surrogate tasks.
- Predict consequences of own actions.
- Learn before task is presented; prepare for the future.

Demonstration Application: Intelligent Visual Seeking

- Our demonstration application: *Intelligent Visual Seeking*
  - Use THOR virtual environment (see images below).
  - Compete in Allen Institute Visual Challenge (AIVC).
  - Leverage **all five core technologies** for superior results.

- Phase I: Focus on core technologies with real images/video.
- Phase II: State-of-the-art on AIVC intelligent visual seeking.
- Technology will have broad impact across core vision, robotics, and machine learning applications.

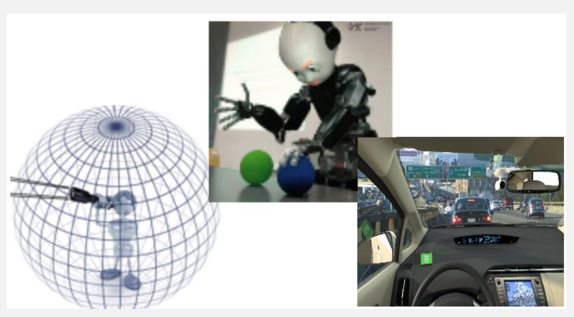
Approach to Lifelong Visual Learning

Status quo:

Learning and inference with "disembodied" snapshots.

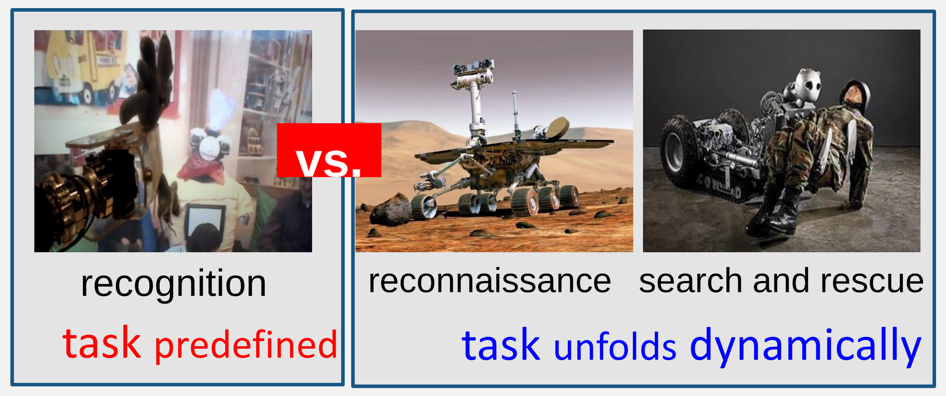
On the horizon:

Visual intelligence in the context of acting and moving in the world.



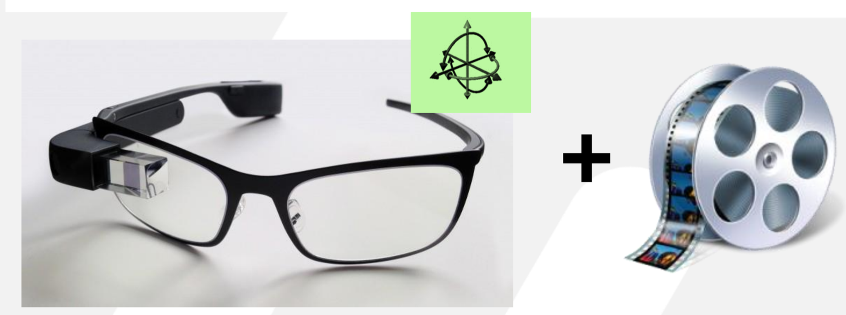
Core research directions: new capabilities

Learning to explore new environments



Embodied visual representations

"how I move" ↔ "how my visual surroundings change"



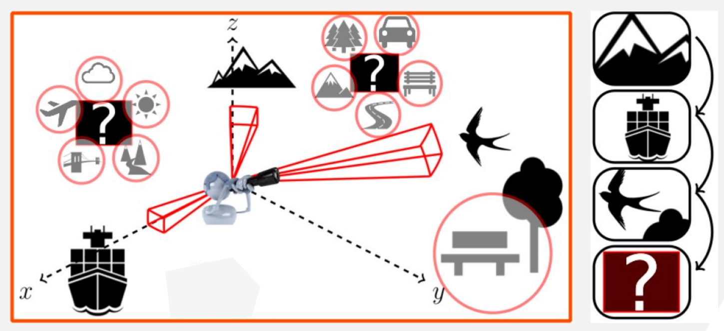
Ego-motion motor signals + Unlabeled video

Adversarial self-play

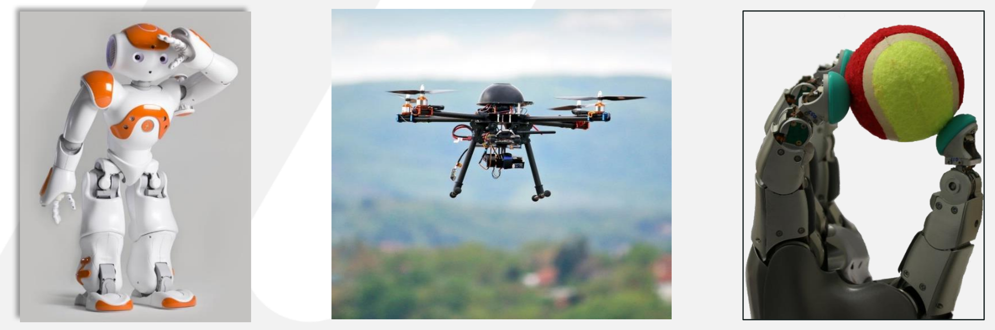
Well defined game rules → Train by self-play

No given game rule → Can't do self-play Can we invent a game rule?

Learning efficient "looking around" policies

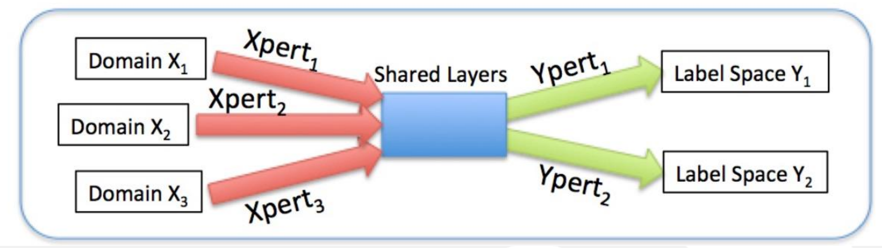


Actively moving to recognize



What motions or manipulations are needed?

Lifelong mixture of experts



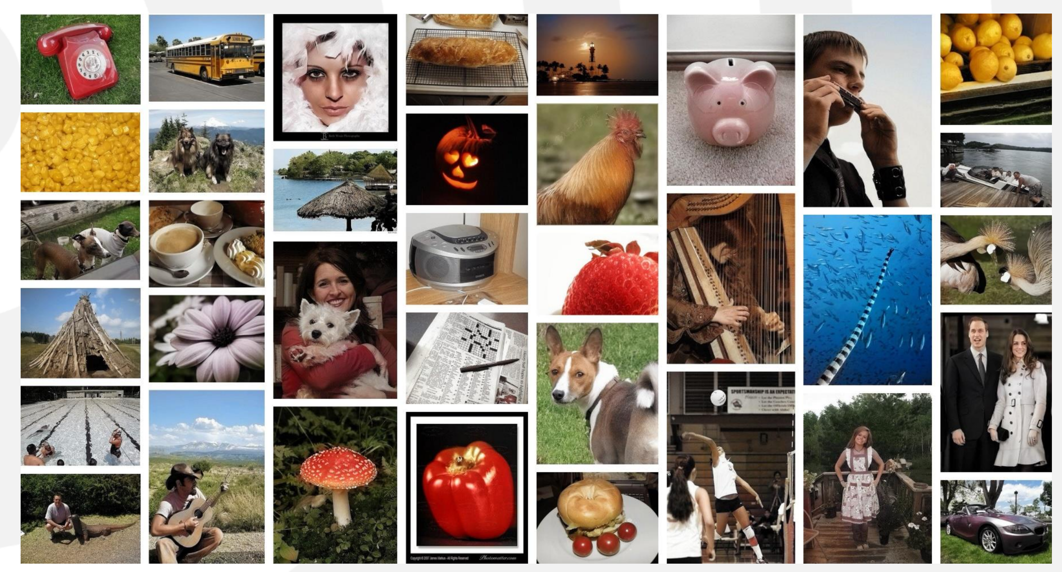
Self-supervision by proxy tasks

- Key intuition: forcing a machine to "acknowledge" structure in the visual world helps learn meaningful representations inside the machine
- Deep learning, where representations are learned end-to-end jointly with the task solver, is particularly suited to this.
- We aim to design proxy tasks that would be tied to such semantically meaningful structure
- "Self-supervision": the task is ostensibly still supervised but the supervisory signal is naturally embedded in the images themselves; no need for designing human-driven labels and annotations.

Learning by colorization

Convolutional NN trained to recover color from gray-scale images; No human-made labels!

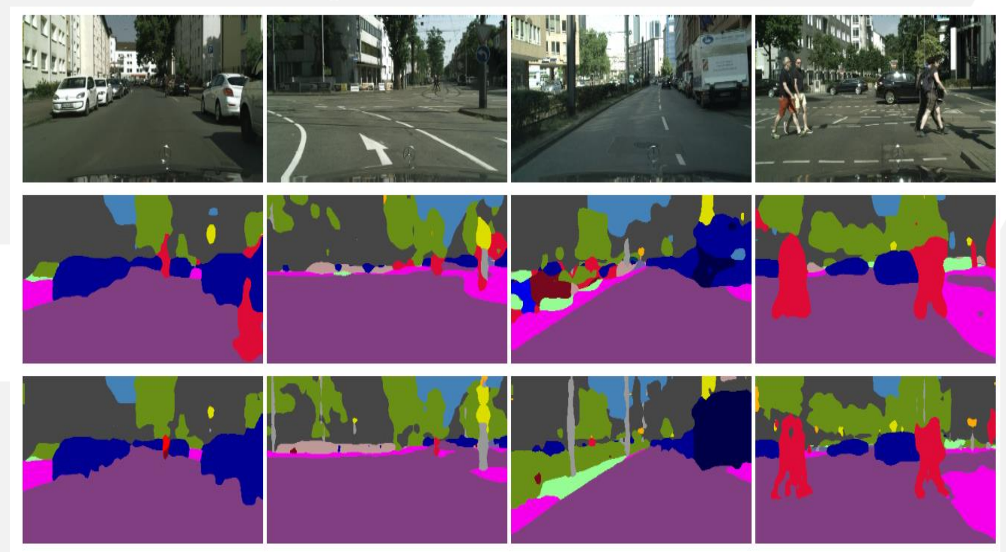
Can use any color images for training



Learning by depth estimation

Intuition: we rely on semantics to judge distance
Compute approximate depth from **motion in video**
Then train a neural network to predict this depth **from a single image.**

Use this **self-supervised pre-trained network** as a starting point for fine-tuning on new tasks like **semantic segmentation** and **improved depth estimation.**



No pre-training

Self-supervised pre-training

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)
The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

THE ELECTRONICS RESURGENCE INITIATIVE